

CLIPPED WORDS

TED CLARKE

Newquay, Cornwall, England

Editor's Note: Ted Clarke here describes a computer program for compressing word lists, reminiscent of the methods used by National Puzzlers' League formists more than a half-century ago, which (unlike other file-compression techniques) can be used without restoring the words first to uncompressed form.

The increasing practice of committing large amounts of text, such as dictionaries and thesauri, to computer-based utilities has produced a number of text-compression techniques for reducing the amount of storage space. There are several industry-standard file-compression utilities in the public domain, any of which may be used to condense the space requirements of text files. My WORDSWORTH files, now containing over 508,000 dictionary words and phrases, were first processed in this fashion, resulting in roughly a 50% saving in disk space. However, files treated in this way need in the majority of cases to be restored to full length in order that the program can read them.

I was attracted to a method which I called 'clipping' because it sheared off the leading group of letters from any succeeding word in a list when these were an exact repetition of those in the preceding word. Originally I wasn't sure that I could make such a compression readable by all of WORDSWORTH's manipulative options. I then came across a Shareware program, WORDFIND, which used this method, giving me the incentive to try to develop clipping into a form which would allow editing from within the program. I also hoped to achieve a screen display which would present all unclipped letters in their true columnar position relative to the other words, such as shown at the left. I was successful in both these objectives.

| | | |
|----------|----------|--|
| CAATINGA | CAATINGA | |
| CABALISM | BALISM | |
| CABALIST | T | |
| CABALLED | LED | |
| CABALLER | R | |
| CABARETS | RETS | |

How much is saved? The original file can be placed in rows, with hyphens indicating the two extra characters for the line-feed and carriage-return: CAATINGA--CABALISM--CABALIST--CABALLED--CABALLER--CABARETS... The clipped list is produced from the sequential file, as a binary file, shown as follows: CAATINGA-BALISM-T-LED-R-RETS...

The compression is obvious. It is further enhanced by discarding the carriage-return character. The line-feed character (or at least some separator) is necessary to define the end of one and the beginning of the next clipped word. A set of 18 words (only the first 6 given above) were clipped by 56.9% by discard-

ing 82 of their 144 letters. If the extraneous characters are taken into account, the overall reduction is from 180 to 80 characters, or bytes. Thus the overall reduction in file length appears to be 55.6%, i.e., slightly less than the figure based on a letters count alone. This is one of those occasions - as with comparisons of inflation indices - where the percentage figure can be confusing. The overall reduction of 100 characters in file length is in fact a further reduction of 18 (100 - 82) on the letters-only figure. This added contribution of one character per word is a bonus, no matter what the percentage figures might indicate.

The saving due to clipping is strongly dependent upon word length, as shown in the table below.

| No. of Letters | Original Files | Clipped Files | SFX Files | Clipped Percent | SFX Percent |
|-------------------|-------------------|------------------|--------------|--------------------|----------------|
| 2 | 1292 | 672 | 16546 | 52.0 | |
| 3 | 8600 | 3814 | 18365 | 44.4 | |
| 4 | 39222 | 15961 | 25766 | 40.7 | |
| 5 | 100149 | 41354 | 40508 | 41.3 | 40.5 |
| 6 | 204112 | 86939 | 65741 | 42.6 | 32.2 |
| 7 | 342882 | 151353 | 99416 | 44.1 | 29.0 |
| 8 | 498980 | 227251 | 137573 | 45.5 | 27.6 |
| 9 | 601755 | 282078 | 166935 | 46.9 | 27.7 |
| 10 | 647736 | 314539 | 181606 | 48.6 | 28.0 |
| 11 | 589394 | 298938 | 173121 | 50.7 | 29.4 |
| 12 | 498400 | 263980 | 154571 | 53.0 | 31.0 |
| 13 | 389190 | 215940 | 130028 | 55.5 | 33.4 |
| 14 | 273488 | 161308 | 103289 | 59.0 | 37.8 |
| 15 | 177327 | 109929 | 77352 | 62.0 | 43.6 |
| 16 | 106704 | 69332 | 56118 | 65.0 | 52.6 |
| 17 | 63327 | 43261 | 42938 | 68.3 | 67.8 |
| 18 | 37680 | 27091 | 34383 | 71.9 | |
| | 4577328 | 2313740 | 1524256 | 50.5 | 33.3 |

The SFX denotes a further compression to a Self-eXtracting File, which will decompress itself to the Clipped File. SFX files carry the instructions for their running and, therefore, contain overheads in excess of the compressed file data. Small sequential files carry a disproportionate amount of such overheads, with the result that there is no overall reduction (see files for words of 2,3,4 and 18 letters). 1524256 bytes is a smidgen too large for a standard 1.44 Mb floppy, but, combined with the program for WORDSWORTH, requires only two disks with ample reserve for expansion of the vocabulary.